Voice analysis rehabilitation platform based on LSTM algorithm

Alessandro Massaro*, Giacomo Meuli,
Nicola Savino and Angelo Maurizio Galiano

Dyrecta Lab Srl,
Via Vescovo Simplicio, N. 45,
70014 Conversano, BA, Italy
Email: alessandro.massaro@dyrecta.com
Email: giacomo.meuli@dyrecta.com
Email: nicola.savino@dyrecta.com
Email: maurizio.galiano@dyrecta.com
*Corresponding author

Abstract: The proposed work discusses the results of a research project based on the recognition of correctly pronounced words and phrases by implementing a web platform implementing an acoustic training model. The acoustic training model is performed by a long short-term memory – LSTM – algorithm, able to recognise the speech disorder by assigning a score for each test type. The paper discusses the platform design and implementation. The tests are performed for different kind of exercises in rehabilitation patterns. The adopted approach is based on the formulation of acoustic model integrating a training dictionary of correct phonemes to pronounce. The platform enables a real time automatic score of the performed exercises and the test planning. The LSTM training dataset can be enriched by adding new exercise to learn. The output graphical dashboards enforce clinical evaluations and reporting.

Keywords: speech disorder recognition; long short-term memory; LSTM; telemedicine platform.

Reference to this paper should be made as follows: Massaro, A., Meuli, G., Savino, N. and Galiano, A.M. (2022) 'Voice analysis rehabilitation platform based on LSTM algorithm', *Int. J. Telemedicine and Clinical Practices*, Vol. 3, No. 4, pp.327–340.

Biographical notes: Alessandro Massaro carried out scientific research at the Polytechnic of Marche, at CNR, and at Italian Institute of Technology (IIT) as Team Leader in nanocomposite sensors for industrial robotics. He is in MIUR register as scientific expert in competitive industrial research and social development, and he is currently head of the Research and Development section and scientific director of MIUR Research Institute Dyrecta Lab Srl. He is a member of the International Committee of Measurers IMEKO and an IEEE senior member. He received an award from the National Council of Engineers as Best Engineer of Italy 2018 (Top Young Engineer 2018).

Giacomo Meuli is a researcher at the Dyrecta Lab Research Institute. He has a degree in Computer Engineering from the Polytechnic of Bari and specialises in information systems in the medical field. His research is currently focused on the analysis of the speech signal and in particular on the automatic evaluation of the pronunciation of words. He has carried out projects in various fields including telemedicine, smart agriculture and forecasting systems based on artificial intelligence algorithms.

Nicola Savino graduated in Electronic Engineering and is a researcher in private research centres in the south of Italy since 2005. He has directed several regional, national and European industrial research projects on ambient intelligence, assisted living, robotics for neuro-rehabilitation, socially assistive robotics, assistive building automation, design and development of complex medical devices, precision farming, energy efficiency systems and smart cities. These projects have generated several scientific publications and international patents. Currently, he is the Operative Director in Dyrecta LAB and coordinates the development of research projects and the product development of integrated platform in biomedical and non-disruptive diagnostic domains.

Angelo Maurizio Galiano is the CEO of Dyrecta Lab Srl – research institute accredited by the Italian Ministry of University and Scientific Research. He has more than 20 years of experience in the field of information technologies. He received his MS degree in Education Science in 2009. His current research interests include neural networks, smart health and predictive analytics.

1 Introduction

Some problems of vowels pronunciation naturally depend on the specific aptitude of citizens to speak in different ways, so it is problematic to be able to speak in another language with the correct pronunciation (Ali, 2013; Hassan, 2014). The analysis of the correct pronunciation is a crucial aspect for different speech disorder pathologies such as stuttering being different the production of syllables: the phonetic configurations determining an increase in stuttering episodes are articulately especially for syllabic configurations, homorganic consonant connections, etc. (Balbo et al., 2012). The study of phonology therefore helps to solve various problems by defining specific sound patterns (Luo, 2014), which are characteristics of each language. In this direction, algorithms of interest are the automatic speech recognition (ASR) algorithms able to recognise a speech in a specific language (Forsberg, 2003). Such tools can be of considerable help, for the creation of a reference 'vocabulary' associated with a specific language (Forsberg, 2003). The creation of the reference vocabulary is the first step for the generation of a basic acoustic model which will be 'enforced' by means of the correct exercise formulation. One of the methods to characterise this variability of speech is the extraction of the cepstral coefficients that represent the spectral envelope of a speech signal frame (Salvi et al., 2011). In any case, the temporal analysis of the vocal signal represents an important element of the voice disorder analysis associated with the pronunciation of certain syllables (Galatà, 2013). Some researchers have classified the voice disorders such as (Omori, 2011):

- vocal cord nodules
- polypoid vocal cords (Reinke's oedema)
- vocal cord atrophy
- sulcus vocalis
- laryngeal granuloma
- functional aphonia

- spasmodic dysphonia
- dysphonia plicae ventricularis
- hypotonic voice disorders
- mutational voice disorders
- essential tremor.

The difficult to understand possible voice disorder causes can be correlated to other physical disorders: as example spasmodic dysphonia may depend on muscle problems (Revelo, 2009), thus increasing the difficulty to find a correct rehabilitation. The temporal and frequency analysis of the voice helps to understand also the health state of a patient (Dejonckere, 2010; Compagnucci et al., 2014). The long short-term memory (LSTM) and in general neural networks (Massaro et al., 2019), could be applied supporting the recognition of the words and phrases thus facilitating the creation and the improvement of the acoustic model. Recently, some researchers focused the attention on the use of deep learning convolutional neural network for the correct pronunciation of Arabic phonemes (Nazir et al., 2019) and for phonetic duration modelling (Wei et al., 2019). In this direction, ASR, improving phonetic deep learning (Pipiras et al., 2019) and neural network architecture (Zhang et al., 2020), is an actual research issue. LSTM network are good candidate for automatic phoneme recognition (Zhang et al., 2016), and represents a flexible tool because can be adopted also for dialect speech recognition (Ying et al., 2019), and for the classification of different forms of stutters (Kourkounakis et al., 2020). The LSTM networks has been also adopted for the detection of Parkinson's disease using subject's voice samples (Rizvi et al., 2020), and for emotion recognition (Wang et al., 2020), and for the dialect identification (Ye et al., 2019), thus confirming the sensitivity of the approach to detect variable pronunciations. The integration of the intelligent algorithms in web-based platforms (Galiano et al., 2016) could facilitate the remote assistance (Massaro et al., 2020), and the improvements of the training models also by using big data repositories (Massaro et al., 2018a, 2018b). Following the topics of the state-of-the-art, the proposed paper discusses some research results of a web-based platform integrating approaches and methodologies suitable for voice disorder recognition. The paper is structured as follows:

- is described the adopted acoustic model based on the adoption of the LSTM method, by describing the training and the testing processes for a rehabilitation exercise
- is discussed the main specifications and the platform design by unified modelling language (UML) implementing use case diagram and class diagram describing the platform functions
- is shown the layout of the frontend interface
- is tested the platform by estimating the score for the pronunciations of different words.

2 Model

In Figure 1 is illustrated a preliminary flowchart of the model concerning creation of the experimental acoustic model describing the following main steps:

1 The model training is performed to extract the features able to create the initial acoustic model; the acoustic model is enriched during the time by new words correctly pronounced constituting the model vocabulary.

2 A first test (test 1) is executed, and the features are extracted and compared with correctly pronounced words and phrases present in the vocabulary.

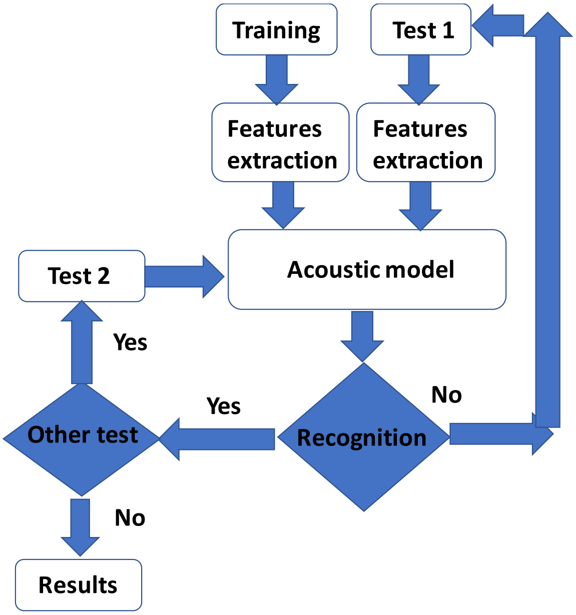
3 Case A: The LSTM algorithm recognises the words (recognition) by allowing to continue with a successive test (test 2).

4 Case B: The test 1 is created to add more cases (new exercises correctly pronounced) into the enriched acoustic model (the vocabulary is enriched in order to perform other exercises including pronounces of phonemes).

5 The platform provides as results graphical output about the patient exercises.

In the proposed paper are discussed the design and the implementation of the platform based on the flowchart of Figure 1 able to recognise voice disorders: the voice features extraction is performed for the testing and for the training model, improving the acoustic model and the training dictionary, the acoustic model recognises errors or correctly pronounced words thus providing a scoring and enabling the possibility to perform in succession other tests.

Figure 1 Basic algorithm of acoustic model used for speech disorder recognition implementing LSTM recognition approach (see online version for colours)



The platform, named ‘voice analysis’, implements the LSTM training and testing flowcharts indicated in Figure 2 and Figure 3, respectively: the training model is able to enrich the training dictionary of the acoustic model by initially adopting a basic dictionary of words correctly pronounced; the testing phase compares input words with the acoustic model data, and training data for the specific exercise to perform by checking the grammar.

Figure 2 Workflow about the LSTM training phase (see online version for colours)

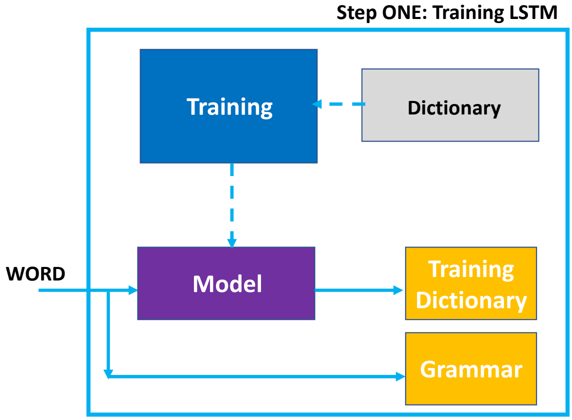
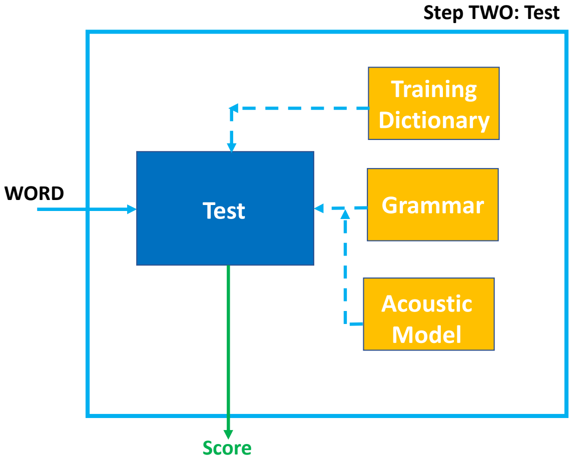


Figure 3 Workflow describing the testing phase (see online version for colours)



3 Voice analysis platform design

3.1 Main platform specifications

The preliminary speech recognition specifications are:

a speech detection and features extraction (features contained in the vocal signals that are important for the phonemes recognition in a word or in a sentence)

b comparison of the extracted features with ones contained in a database (acoustic model able to identify the correct pronounced word)

c training module creating the acoustic model

d possibility of inserting new exercises for a vocal synthesis of pronunciation of the word or for the analysis of single phonemes or syllables

e failure analyser (if the patient fails to correctly pronounce the written or listened text, the program signals it by assigning a low score to each pronounced word)

f timing for the exercise pronunciation

g possibility of setting pronunciation speed time

h possibility to include graphical elements (Pirovano et al., 2016) to follow the exercises to execute (the platform shows graphical objects simulating the individual exercises as games).

3.2 Spontaneous eloquium specifications

The spontaneous eloquium exercise of the proposed platform, requires that the user speaks freely in the microphone observing in a text window the speech. The correctness or otherwise of the goodness of the pronunciation is given by the comparison with the words inserted in the ‘vocabulary’ (dictionary) of the acoustic model of the software.

Below are examples of sounds that acoustic model uses for the comparison:

- *Ci-gi* affricate phones
- *f-v* fricative phones
- *m-n* nasal phones
- *gn-gl* nasal phone and liquid phone
- *k-g* occlusive phones
- *t-d* occlusive phones
- *p-b* occlusive phones
- *s* fricative sound
- *sc* fricative phone
- *z* affricate phone
- *L* liquid phone
- *r* vibrating phone.

3.3 UML design

The above specifications of the ‘voice analysis’ platform are expressed by the UML diagrams. UML is a graphical language recently adopted for the design of healthcare platforms (Khalid et al., 2019; Variani et al., 2017). In Figure 4 is illustrated the UML case diagram indicating the following system actors:

- patient connecting to the platform
- registered user (enabling patients for particular platform accesses)
- administrator (supervisor user enabling patient connection).

Figure 4 Voice analysis platform: UML use case diagram

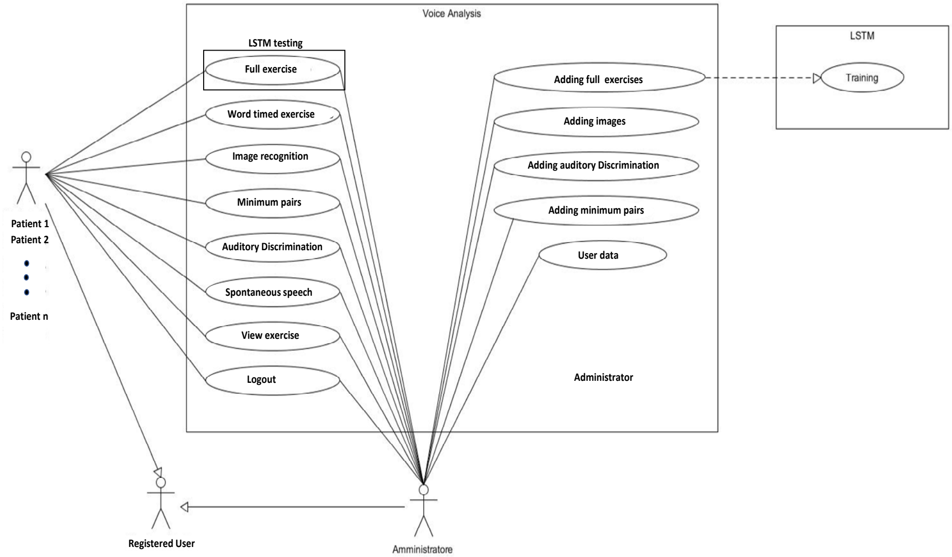


Figure 4 shows that LSTM network works for model training and for exercise testing. Moreover, the platform could also integrate image processing functions able to recognise the patient and to read facial expressions. The prototype platform will learn by the LSTM training dataset constructed by initially implementing a correct vocal vocabulary model in Italian language (acoustic model of Figure 1). In the UML class diagram of Figure 5 are listed all the implemented functions of the prototype platform where the main class is the phoneme recognition (fonema): all classes are linked to structure a full exercise including phrases, tongue twisters exercises, and complex words to pronounce.

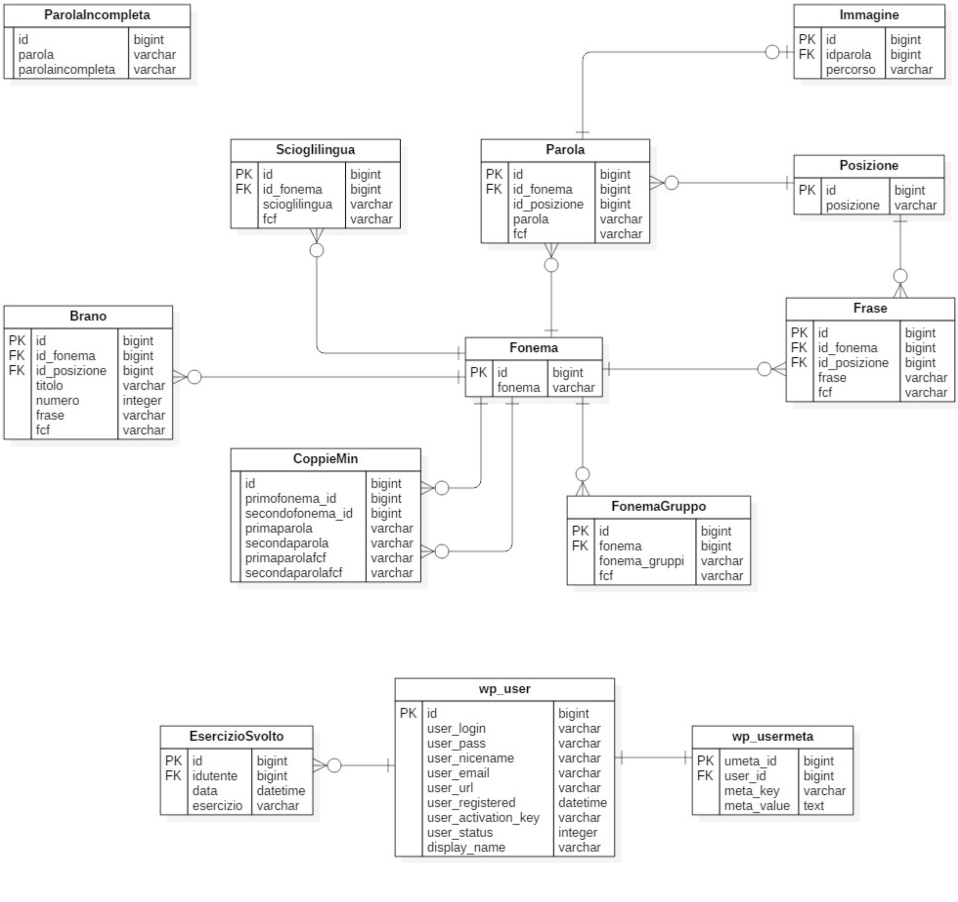
The users (patients) which are registered can participate to test exercises for:

- word pronunciation
- word position
- graphic guided wizard exercise
- phoneme group recognition (FonemaGruppo)

- tongue twisters exercise (Scioglilingua)
- passage exercise (Brano)
- incomplete word recognition (ParolaIncompleta).

The training model is based on the LSTM approach which is a recurrent neural network (RNN) architecture (Bianchi et al., 2017) used in the field of deep learning.

Figure 5 UML class diagram



The recurring networks also provide connections backwards or towards the same level, that is, at each sequence step the cell receives in addition to the input $x(t)$ also its output from the previous step $y(t-1)$. This allows the network to base its decisions on the past history (memory effect) or on all the elements of sequence and on their mutual position.

A cell is a part of the recurring network that preserves an internal state $h(t)$ for each instant of time, which depends on the input $x(t)$ and the previous state $h(t-1)$.

$$h(t) = f\left(h(t-1), x(t)\right) \tag{1}$$

Cells have difficulty remembering the inputs of distant steps and therefore memory tends to fade. To solve this problem, more LSTM cells are used.

In LSTM, the state is divided into two vectors:

$h(t)$ is the short-term state

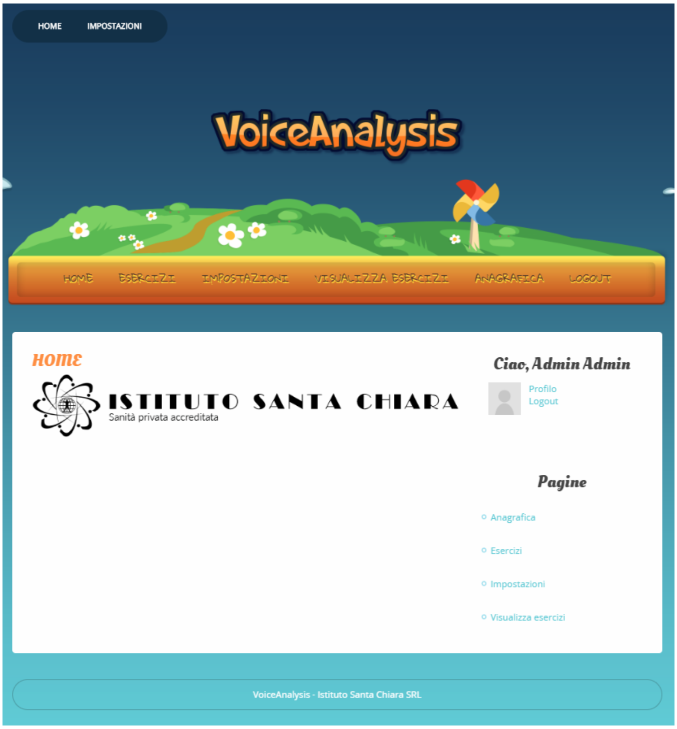
$c(t)$ is the long-term state.

In the learning process (unfolding), the cell must consider what information to keep (forgot gate) of the past state $c(t-1)$ and what to extract and add (input gate) from the current input $x(t)$. Instead, to calculate the output $y(t)$, combine the current input (output gate) with the information extracted from long-term memory (Massaro et al., 2019).

4 Platform development and testing

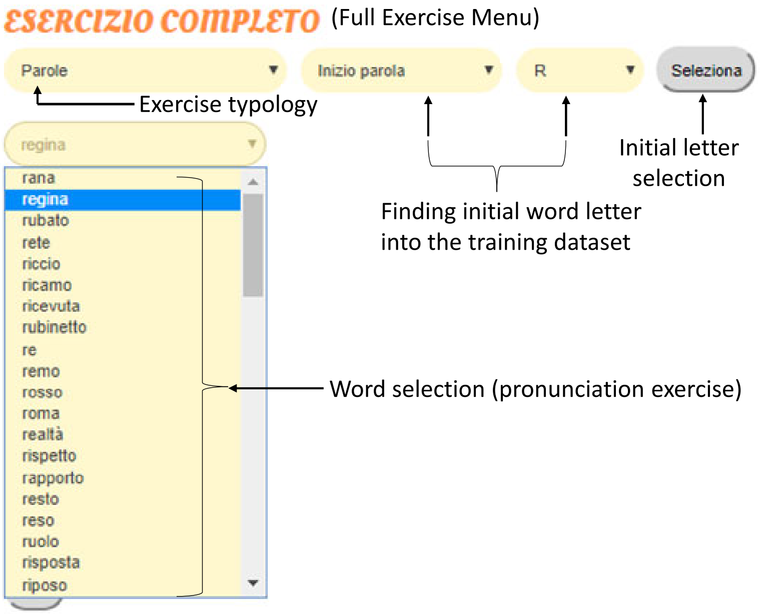
The platform has been developed by adopting python and PHP scripts (webservice) for backend development and javascript and ajax call for frontend. Data are processed by a local processor (server machine). The used LSTM model is based sequence to sequence approach (seq2seq) which is a general-purpose encoder-decoder framework (Michael et al., 2019) for Tensorflow (Sanchez et al., 2020). In Figure 6 is illustrated the graphical user interface (GUI) of the prototypal platform of the main webpage (homepage). By means of the GUI, the registered user can access to the platform by executing the indicated exercises. By the webpage of Figure 6, it is possible to entry in the administration area, to select exercise by linking the patient, and to visualise the scoring and complete list of the available exercises.

Figure 6 GUI of the 'voice analysis' platform (see online version for colours)



In Figure 7 is shown an example of the graphical interface of a full exercise selection concerning a word combination to pronounce: the window enables the selection of the word by filtering the initial letters and by showing the different possibilities.

Figure 7 GUI of the ‘voice analysis’ platform: window selecting word (see online version for colours)



(Full Exercise Menu)

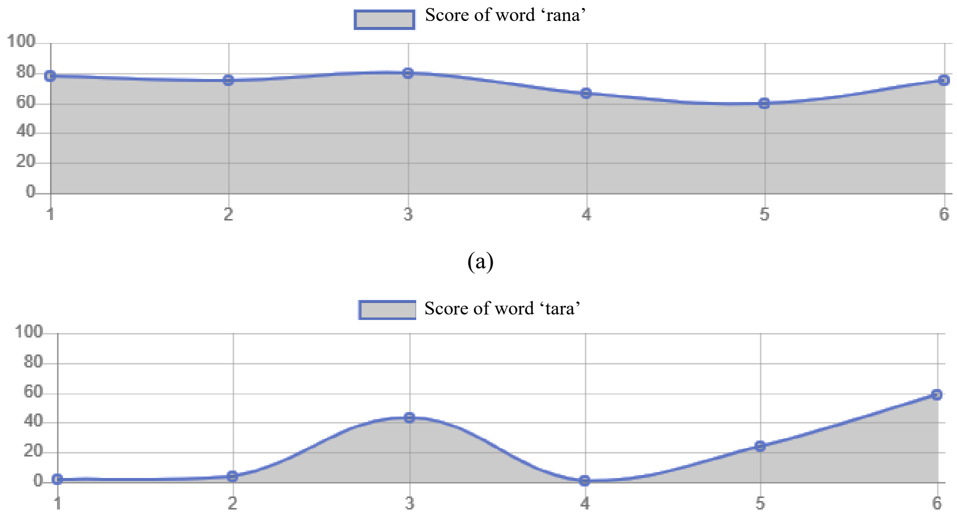
Exercise typology

Initial letter selection

Finding initial word letter into the training dataset

Word selection (pronunciation exercise)

Figure 8 Example of exercise scoring, (a) scoring of word ‘rana’ (b) scoring of word ‘tara’ (see online version for colours)



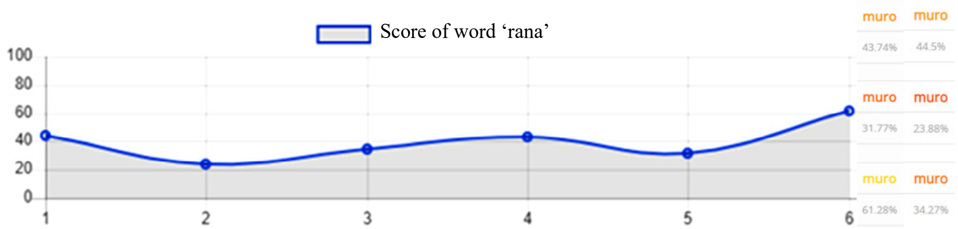
(a)

(b)

The ‘voice analysis’ system attributes a score for each pronounced word. In Figure 8 are related two scoring process of two different words: in the analysed case the first word ‘rana’ is pronounced with a high score, besides the word ‘tara’ is characterised by a low score. Figure 8 represent the dashboard output of the platform supporting clinical evaluations. By improving the training dataset and by implementing different techniques, it is possible to achieve a word error rate (WER) of the order of 6% (Prabhavalkar et al., 2017).

The test outputs can be visualised by illustrating each score with a colour scale bar (see Figure 9), where the red colour indicates a critical word pronunciation: the colour supports the expertise to have in real time the trend of the scoring by supporting the choice of the next exercise to execute. The threshold values of a correct pronunciation can be set in function of the exercise complexity.

Figure 9 Example of exercise scoring by adopting a colour scale bar (pronunciation of the word ‘muro’) (see online version for colours)



5 Conclusions

The paper discusses some results of a project concerning the study and the development of a web platform for voice analysis rehabilitation based on the application of LSTM algorithm useful to recognise the correct vocal pronunciation by enriching the model and the vocabulary with new exercise during the time. The developed prototype ‘voice analysis’ platform provides a guided GUI for the execution of the user exercises and a graphical score. The paper is focused on the description of the adopted technologies by commenting the platform design. The LSTM algorithms can be potentially applied for other kinds of acoustic models of different languages. The discussed results enhance the possibility to realise a web-based backend and frontend system automatising the exercises to perform: the UML design and the block diagrams shows how are integrated the different platform functions and classes by explaining the recognition mechanisms. Future directions of the proposed results are in the automatic classification of voice disorders by taking into account dialectical cadences. The platform will be optimised in order to check the care evolution and to propose automatically new exercises in function of the scores acquired of each patient. In this way will be possible also to formulate, by means of a decision supporting system, a dynamic rehabilitation patterns thus optimising care process time. A further improvement could be achieved by applying vocal spectroscopy (Campanella et al., 2019; Song et al., 2017) and automated exercise by the web service. The innovative approach to use of the ‘remote and automated rehabilitation’ provides a new concept of telemedicine based on the goal to optimise human resources and to increase patient security especially in pandemic periods.

Acknowledgements

The work has been developed in the framework of the project: 'Programma computerizzato per il trattamento dei disturbi del linguaggio: 'VoiceAnalysis'' ('Computerized program for the treatment of language disorders 'VoiceAnalysis''). The authors would like to thank the researchers V. Calati, D. Carella, A. Colonna, R. Cosmo, V. Custodero, G. Fanelli, A. Galiano, R. Guglielmi, M. Legrottaglie, A. Leogrande, A. Lombardi, G. Lonigro, A. Lorusso, L. Maffei, N. Malfettone, S.F. Francesco, L. Pellicani, R. Porfido, N. Savino, D. Suma, F. Tarulli and E. Valenzano.

References

Ali, E.M.T. (2013) 'Pronunciation problems: acoustic analysis of the English vowels produced by Sudanese learners of English', *International Journal of English and Literature*, Vol. 4, No. 10, pp.495–507, DOI: 10.5897/IJEL12.031.

Balbo, D., Verdurand, M., Rossato, S. and Zmarich, C. (2012) 'La produzione di sillabe nella balbuzie in condizioni di feedback uditivo normale e alterato', in *Proceeding of International Conference on Stuttering*, Omega Edizioni, Torino, Roma, 7–9 June, pp.177–188.

Bianchi, F.M., Maiorino, E. and Kampffmeyer, M.C. (2017) 'Recurrent neural network architectures', in *Recurrent Neural Networks for Short-term Load Forecasting* [online] http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-319-70338-1_3.

Campanella, A., Manca, F., Marin, C., Bosna, V., Salonnna, I., Galatola, M. and Sabella, E.A. (2019) 'Augev method and an innovative use of vocal spectroscopy in evaluating and monitoring the rehabilitation path of subjects showing severe communication pathologies', *International Journal of Clinical Medicine*, Vol. 10, pp.27–52, DOI: 10.4236/ijcm.2019.102004.

Compagnucci, M., Mazzocchi, R. and Centorrino, S. (2014) 'Analisi oggettiva della voce in ambulatorio logopedico: uso di uno strumento di analisi vocale', *Logopaedia*, Vol. 12, No. 1, pp.21–31 [online] https://www.fli-lazio.it/images/allegati/rivista/logopedia_imp_1_14_low2.pdf.

Dejonckere, P.H. (2010) 'Assessment of voice and respiratory function', in Remacle, M. and Eckel, H.E. (Eds.): *Surgery of Larynx and Trachea*, Springer-Verlag, Berlin, Heidelberg, DOI: 10.1007/978-3-540-79136-2_2.

Forsberg, M. (2003) *Why is Speech Recognition Difficult?* [online] https://www.researchgate.net/publication/228763868_Why_is_speech_recognition_difficult (accessed 1 October 2020).

Galatà, V. (2013) 'Multimodalità e multilingualità: la sfida più avanzata della comunicazione orale' *Proceeding of 9° Convegno Nazionale AISV*, 21–23 January.

Galiano, A., Massaro, A., Boussahel, B., Barbuzzi, D., Tarulli, F., Pellicani, L., Renna, L., Guarini, A., De Tullio, G., Nardelli, G., Bonaduce, R., Minoia, C., Ciavarella, S., De Fazio, V., Negri, A. and Marchionna, C. (2016) 'Improvements in haematology for home health assistance and monitoring by a web based communication system', *Proceeding of IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, DOI: 10.1109/MeMeA.2016.7533762.

Hassan, E.M.I. (2014) 'Pronunciation problems: a case study of English language students at Sudan University of Science and Technology', *English Language and Literature Studies*, Vol. 4, No. 4, pp.31–44, DOI: 10.5539/ells.v4n4p31.

Khalid, M., Afzaal, H., Hassan, S., Zafar, N.A., Latif, S. and Rehman, A. (2019) 'Automated UML-based formal model of e-health system', *IEEE Proceeding of 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, DOI: 10.1109/MACS48846.2019.9024830.

Kourkounakis, T., Hajavi, A. and Etemad, A. (2020) 'Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory', *IEEE Proceeding International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, DOI: 10.1109/ICASSP40776.2020.9053893.

Luo, J. (2014) 'A study of mother tongue interference in pronunciation of college English learning in China', *Theory and Practice in Language Studies*, Vol. 4, No. 8, pp.1702–1706, DOI: 10.4304/tpls.4.8.1702-1706.

Massaro, A., Galiano, A., Scarafile, D., Vacca, F., Vacca, A., Frassanito, A., Melaccio, A., Solimando, A., Ria, R., Calamita, G., Bonomo, M., Vacca, F., Gallone, A. and Attivissimo F. (2020) 'Telemedicine DSS-AI multi level platform for monoclonal gammopathy assistance', *IEEE Proceeding of MeMeA 2020*, ISBN: 978-1-7281-5386-5.

Massaro, A., Maritati, V., Giannone, D., Convertini, D. and Galiano, A. (2019) 'LSTM DSS automatism and dataset optimization for diabetes prediction', *Applied Sciences*, Vol. 9, No. 17, pp.1–22, DOI: 10.3390/app9173532.

Massaro, A., Maritati, V., Savino, N. and Galiano, A. (2018a) 'Neural networks for automated smart health platforms oriented on heart predictive diagnostic big data systems', *IEEE Proceeding AEIT 2018*, DOI: 10.23919/AEIT.2018.8577362.

Massaro, A., Maritati, V., Savino, N., Galiano, A., Convertini, D., De Fonte, E. and Di Muro, M. (2018b) 'A study of a health resources management platform integrating neural networks and DSS telemedicine for homecare assistance', *Information*, Vol. 9, No. 176, pp.1–20 [online] https://doi.org/10.3390/info9070176.

Michael, J., Labahn, R., Gruning, T. and Zollner, J. (2019) *Evaluating Sequence-to-sequence Models for Handwritten Text Recognition*, arXiv: 1903.07377v2.

Nazir, F., Majeed, M.N., Ghazanfar, M.A. and Masqsood, M. (2019) 'Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes', *IEEE Access*, Vol. 7, No. 1, pp.52589–52608, DOI: 10.1109/ACCESS.2019.2912648.

Omori, K. (2011) 'Diagnosis of voice disorders', *JMAJ*, Vol. 54, No. 4, pp.248–253.

Pipiras, L., Maskeliunas, R. and Damaševicius, R. (2019) 'Lithuanian speech recognition using purely phonetic deep learning', *Computers*, Vol. 8, No. 76, pp.1–15, DOI: 10.3390/computers8040076.

Pirovano, M., Surer, E., Mainetti, R., Lanzi, P.L. and Borghese, N.A. (2016) 'Exergaming and rehabilitation: a methodology for the design of effective and safe therapeutic exergames', *Entertainment Computing*, Vol. 4, pp.55–65 [online] https://doi.org/10.1016/j.entcom.2015.10.002.

Prabhavalkar, R., Sainath, T.N., Wu, Y., Nguyen, P., Chen, Z., Chiu, C-C. and Kannan, A. (2017) *Minimum Word Error Rate Training for Attention-based Sequence-to-sequence Models*, arXiv: 1712.01818v1.

Revelo, O. (2009) 'Spasmodic dysphonia: evaluation and management', *Grand Rounds Presentation*, Department of Otolaryngology, UTMB.

Rizvi, D.R., Nissar, I., Masood, S., Ahmed, M. and Ahmad, F. (2020) 'An LSTM based deep learning model voice-based detection of Parkinson's disease', *International Journal of Advanced Science and Technology*, Vol. 29, No. 5s, pp.337–343.

Salvi, G., Tesser, F., Zovato, E. and Cosi, P. (2011) 'Analisi gerarchica degli inviluppi spettrali differenziali di una voce emotiva', *Proceeding of 7° Convegno dell'Associazione Italiana Scienze della Voce*, pp.369–379.

Sanchez, S.A., Romero, H.J. and Morales, A.D. (2020) 'A review: comparison of performance metrics of pretrained models for object detection using the TensorFlow framework', *Materials Science and Engineering*, Vol. 844, pp.1–15, DOI: 10.1088/1757-899X/844/1/012024.

Song, E., Soong, F.K. and Kang, H-G. (2017) 'Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 11, pp.2152–2161, DOI: 10.1109/TASLP.2017.2746264.

Variani, E., Bagby, T., McDermott, E. and Bacchiani, M. (2017) 'End-to-end training of acoustic models for large vocabulary continuous speech recognition with Tensorflow', *Proceeding of Iterspeech 2017*, DOI: 10.21437/Interspeech.2017-1284.

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J. and Tarokh, V. (2020) *Speech Emotion Recognition with Dual-sequence LSTM Architecture*, arXiv: 1910.08874v4.

Wei, X., Hunt, M. and Skilling, A. (2019) *Neural Network-based Modeling of Phonetic Durations*, arXiv: 1909.03030v1.

Ye, S., Li, C., Zhao, R. and Wu, W. (2019) 'NOAA-LSTM: a new method of dialect identification', in *Artificial Intelligence and Security*, Springer, Cham [online] https://doi.org/10.1007/978-3-030-24274-9_2.

Ying, W., Zhan L. and Deng, H. (2019) 'Sichuan dialect speech recognition with deep LSTM network', *Frontiers of Computer Science*, Vol. 14, No. 1, pp.378–387 [online] https://doi.org/10.1007/s11704-018-8030-z.

Zhang, B., Gan, Y., Song, Y. and Tang, B. (2016) 'Application of pronunciation knowledge on phoneme recognition by LSTM neural network', *IEEE Proceeding of 23rd International Conference on Pattern Recognition (ICPR)*, DOI: 10.1109/ICPR.2016.7900078.

Zhang, L., Zhao, Z., Ma, C., Shan, L., Jiang, L., Deng, S. and Gao, C. (2020) 'End-to-end automatic pronunciation error detection based on improved hybrid CTC/attention architecture', *Sensors*, Vol. 20, No. 7, pp.1–24, DOI: 10.3390/s20071809.